DOCUMENT RESUME

ED 467 821 TM 034 365

AUTHOR	Nandakumar, Ratna; Roussos, Louis
TITLE	CATSIB: A Modified SIBTEST Procedure To Detect Differential Item Functioning in Computerized Adaptive Tests. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Princeton, NJ.
REPORT NO	LSAC-R-97-11
PUB DATE	2001-09-00
NOTE	18p.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	Ability; *Adaptive Testing; *Computer Assisted Testing; Identification; Item Banks; *Item Bias; Item Response Theory; Regression (Statistics); Simulation
IDENTIFIERS	*CATSIB Computer Program; SIBTEST (Computer Program); Type I Errors

ABSTRACT

Computerized adaptive tests (CATs) pose major obstacles to the traditional assessment of differential item functioning (DIF). This paper proposes a modification of the SIBTEST DIF procedure for CATs, called CATSIB. CATSIB matches test takers on estimated ability based on unidimensional item response theory. To control for impact-induced Type I error inflation, the SIBTEST regression correction is shown to have an easily implemented and theoretically justified counterpart in the CAT setting. A simulation study was conducted to evaluate the performance of CATSIB. Simulated test takers were adaptively administered 25 simulated operational items from a pool of 1,000 and were linearly administered 16 simulated pretest items that were evaluated for DIF. The pretest items were designed to represent varying levels of discrimination, difficulty, and amounts of DIF. Sample size varied from 250 to 500 in each group. Simulated levels of impact ranged from 0 to 1 standard deviations difference in mean ability levels. Results show that CATSIB with the regression correction displays impact-induced Type I error inflation. In terms of power, even with as few as 250 test takers in each group, CATSIB had detection rates of 64% or greater for large values of DIF. When sample size was increased to 500 in each group, these power rates increased to more than 90%. CATSIB displayed nearly unbiased estimation under nearly all the simulated conditions. (Contains 7 tables, 1 figure, and 11 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 467 821

LSAC RESEARCH REPORT SERIES

TM

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. VASELECK

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

CATSIB: A Modified SIBTEST Procedure to Detect Differential Item Functioning in Computerized Adaptive Tests

Ratna Nandakumar
University of Delaware

and

Louis Roussos
Law School Admission Council

Law School Admission Council
Computerized Testing Report 97-11
September 2001

TM034365

ERIC

2

BEST COPY AVAILABLE

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 199 law schools in the United States and Canada.
Copyright © 2001 by Law School Admission Council, Inc.

All rights reserved. No part of this report may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission of the publisher. For information, write: Communications, Law School Admission Council, Box 40, 661 Penn Street, Newtown, PA 18940-0040.

LSAT® and the Law Services logo are registered marks of the Law School Admission Council, Inc.
This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the authors and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

The Law School Admission Council (LSAC) is in the midst of a research program to determine the advisability and feasibility of developing a computerized version of the Law School Admission Test (LSAT). In any testing situation, it is essential that items are fair to all subgroups of test takers. If one subgroup is performing better than another on an item—although both subgroups have been matched on ability—then such an item will be cause for concern. This phenomenon is known as differential item functioning or simply DIF. Even though reliable statistical procedures have been developed for detecting DIF items on paper-and-pencil tests, computerized adaptive tests (CATs) pose obstacles that require the development of new procedures.

DIF analyses require comparing the performance of test takers from different subgroups by carefully selecting test takers who have been matched on some measure of ability level. With a paper-and-pencil test, the matching criterion is typically the number-right score on all the items, including the item being studied for DIF. The inclusion of the score on the studied item helps control for statistical error due to *impact* (group-average ability differences). However, with a CAT different test takers take different items according to their measured ability levels. Thus, they all get similar number-right scores, and number-right score cannot be used as a criterion for ability level matching. Hence, a new matching criterion must be developed. Also, a new method to control for statistical error due to impact must be developed.

The current paper proposes a new DIF procedure for CATs which overcomes these obstacles. The procedure is called CATSIB as it is a modification of the SIBTEST DIF procedure that is used with paper-and-pencil tests. CATSIB matches test takers on estimated ability level, an estimate that a CAT produces for each test taker. To control for statistical error due to impact, a correction is applied to these ability estimates—a correction that is based on a similar correction used with SIBTEST.

To evaluate the performance of the new procedure, a simulation study was conducted. The simulated testing situation consisted of test takers receiving 25 adaptively administered operational items (from a pool of 1,000) and 16 linearly administered pretest items that were evaluated for DIF. The simulated pretest items were statistically designed to display varying known amounts of DIF, and CATSIB was applied to these data to see how well it could detect and estimate these known amounts of DIF. Also, various levels of impact were simulated to see how well CATSIB could control for impact-induced statistical error. The simulation results showed that CATSIB was very effective in controlling statistical error due to impact, even for large average group ability level differences. CATSIB also performed well in detecting and estimating the simulated amounts of DIF in the items, exhibiting detection rates of over 90% for sample sizes of 500 in each subgroup and over 60% for sample sizes of 250 in each subgroup. Future research is planned to further improve CATSIB performance.

Abstract

Computerized adaptive tests (CATs) pose major obstacles to the traditional assessment of differential item functioning (DIF). Test takers cannot be matched on number-right score, and new methods need to be developed to control for Type I error inflation due to mean performance differences (impact) between the reference and focal groups. To this end, a modified SIBTEST procedure—CATSIB—is proposed, which matches test takers on estimated ability based on unidimensional item response theory. To control for impact-induced Type I error inflation, the SIBTEST regression correction is shown to have an easily implemented and theoretically justified counterpart in the CAT setting.

A simulation study was conducted to evaluate the performance of CATSIB. Simulated test takers were adaptively administered 25 simulated operational items from a pool of 1,000, and were linearly administered 16 simulated pretest items that were evaluated for DIF. The pretest items were designed to represent varying levels of discrimination, difficulty, and amounts of DIF. Sample size varied from 250 to 500 in each group. Simulated levels of impact ranged from 0 to 1 standard deviations difference in mean ability levels.

The results showed that CATSIB with the regression correction displayed very good control over Type I error, whereas CATSIB without the regression correction displayed impact-induced Type I error inflation. In terms of power, even with as few as 250 test takers in each group, CATSIB had detection rates of 64% or greater for large values of DIF. When sample size was increased to 500 in each group, these power rates increased to over 90%. CATSIB displayed nearly unbiased estimation under nearly all the simulated conditions.

Introduction

Computerized testing and computerized adaptive testing (CAT) are becoming increasingly popular for standardized tests. For example, test takers can now opt for computerized adaptive tests of the Graduate Management Admission Test (GMAT), Graduate Record Examination (GRE), and the Armed Services Vocational Aptitude Battery (ASVAB). Moreover, the Law School Admission Council (LSAC) is in the midst of a research plan to evaluate the advisability and feasibility of a computerized version of the Law School Admission Test (LSAT). In any testing situation, it is essential that items are fair to all subgroups of test takers. In other words, if—on an item—one subgroup is doing better than the other, although both subgroups have been matched on ability, then such an item will be cause for concern. This phenomenon is known as differential item functioning (DIF). A number of studies have been conducted using different methodologies to investigate DIF on paper-and-pencil tests. In a typical paper-and-pencil test, most DIF analyses involve matching test takers on the number-correct score, a traditional matching criterion. However, in the context of CAT, not all test takers take the same items, and sometimes not all take the same number of items. Hence one cannot directly apply methods developed for paper-and-pencil tests to assess DIF of CAT items. Zwick, Thayer, and Wingersky (1994, 1995) have studied DIF on computerized adaptive test items using a modified version of the Mantel Haenszel methodology. Roussos (1996) has developed a modified version of SIBTEST, called CATSIB, to identify DIF items of an adaptive test using a matching criterion based on CAT ability estimates. Roussos, however, investigated only the Type I error performance of CATSIB and only under standard paper-and-pencil (linear) testing conditions. The purpose of the present study was to investigate the performance of CATSIB to identify DIF items of adaptive tests through simulations. Both the Type I error rate and power level were of interest in the present study. CATSIB will be briefly described below, followed by details of the simulation study, results, and discussion.

CATSIB

An item is said to display DIF when test takers of equal proficiency (on the construct measured by the test), but from separate populations, differ in their probability of answering the item correctly. The proficiency of a test taker on the construct will be referred to as θ. The item that is being tested for DIF is commonly referred to as the *studied item*. The populations of interest for DIF analyses are most commonly based on ethnicity or gender. The populations are categorized into a *reference (R) group* population (for example, Caucasians or males) and a set of *focal (F) group* populations (for example, various minority groups or females).

Using the above terminology, a studied item is said to display DIF when reference group test takers and focal group test takers who have been matched on θ do not have the same probability of a correct response on the item. One common procedure for modeling and simulating DIF in an item is to use a different item response function (IRF) for the reference group than for the focal group. The reference group IRF is denoted by $P_R(\theta)$ and the focal group IRF by $P_F(\theta)$.

Let DIF(θ) be defined as the magnitude of DIF in a studied item at a particular value of θ. A variety of formulations of DIF(θ) exist in terms of $P_R(\theta)$ and $P_F(\theta)$. We employ the formulation used by the Shealy and Stout (1993) SIBTEST procedure,

$$\mathrm{DIF}(\theta) = P_R(\theta) - P_F(\theta), \tag{1}$$

which is also the formulation used by the standardization procedure of Dorans and Kulick (1986). DIF is then defined as an average of DIF(θ) over θ. Employing the SIBTEST terminology, this average is denoted by β, which is given by

$$\beta = \int DIF(\theta) f(\theta) d\theta, \tag{2}$$

where $f(\theta)$ is an appropriate density function on θ such as that for the combined R and F groups. Hence, the null hypothesis for DIF hypothesis testing is stated as

$$H_0 : \beta = 0.$$

An important aspect of any DIF procedure is to select R and F test takers that are matched on ability before comparing their performance on the studied item. For didactic purposes, let us consider the case of estimating β for a pretest studied item. With paper-and-pencil tests, SIBTEST matches test takers based on estimated true score from the operational items of the test. The use of estimated score instead of simple observed score is referred to in the SIBTEST literature as the regression correction. This term refers to the fact that when R and F differ in mean observed score, the regression of true score on observed score will also differ between the two groups ($E_R[T \mid X] \neq E_F[T \mid X]$). Since true score is monotonically related to θ, if this disparity were not corrected for by the regression correction, estimation bias and correspondingly inflated Type I error would result.

Just as the R and F groups may differ in their observed score distribution means in the paper-and-pencil setting, they may differ in their mean values of estimated θ in the CAT setting. In this paper, we will refer to the difference in the means of the θ distributions as *impact*, which we will denote by $d_T = \mu_{\theta_R} - \mu_{\theta_F}$, where μ_{θ_R} and μ_{θ_F} are the means of the ability distributions of the R and F groups, respectively. Just as observed-score impact can result in estimation bias and inflated Type I error in the paper-and-pencil setting, the presence of impact in the CAT setting can have the same effect. Specifically, when impact is present, the use of estimated θ (denoted by $\hat{\theta}$) as the matching variable in the estimation of β could lead to a high Type I error because $E_R[\theta \mid \hat{\theta}]$ can be much different from $E_F[\theta \mid \hat{\theta}]$. Therefore, in order to avoid the high Type I error rate, CATSIB employs a regression correction that is theoretically equivalent to that which SIBTEST employs with paper-and-pencil tests. Instead of matching test takers on $\hat{\theta}$, CATSIB matches test takers on an estimate of $E_G[\theta \mid \hat{\theta}]$, which we denote by $\hat{\theta}^*$. The subscript G on the expectation stands for group membership and indicates that the regression correction is carried out separately for $G = R$ and $G = F$.

The derivation of the formula for $E_G[\theta \mid \hat{\theta}]$ parallels that for $E_G[T \mid X]$ in classical test theory, which is the regression correction formula in Shealy and Stout's (1993) SIBTEST. For convenience, we drop the subscript G notation during the derivation with the understanding that all means, variances, and covariances are in reality carried out separately in the two groups ($G = R, F$). First, we assume that $\hat{\theta}$ is an approximately unbiased estimator for θ, such that

$$\hat{\theta} \approx \theta + e, \tag{3}$$

where e stands for measurement error, which is assumed to have a mean of 0 and a variance of σ_e^2. It is also assumed that e is uncorrelated with θ.

Thus,

$$Var(\hat{\theta}) \approx Var(\theta) + Var(e).$$

Rewriting, we get

$$\sigma_{\hat{\theta}}^2 \approx \sigma_\theta^2 + \sigma_e^2. \tag{4}$$

We also need the equation for $cov(\hat{\theta}, \theta)$, which is given by,

$$cov(\hat{\theta}, \theta) = cov(\theta + e, \theta) = \sigma_\theta^2. \tag{5}$$

Now, we can derive the equation for $\rho_{\hat{\theta},\theta}$, the correlation between $\hat{\theta}$ and θ, which is given by

$$\rho_{\hat{\theta},\theta} = \frac{\mathrm{cov}(\hat{\theta}, \theta)}{\sigma_{\hat{\theta}} \sigma_\theta}.$$

After some algebra we finally get

$$\rho_{\hat{\theta},\theta} = \sqrt{1 - \frac{\sigma_e^2}{\sigma_{\hat{\theta}}^2}} = \frac{\sigma_\theta}{\sigma_{\hat{\theta}}}. \tag{6}$$

By applying standard linear regression theory we obtain the following equation for $E[\theta|\hat{\theta}]$,

$$E[\theta|\hat{\theta}] = E[\theta] + \frac{\rho_{\theta,\hat{\theta}}\sigma_{\theta}}{\sigma_{\hat{\theta}}}(\hat{\theta} - E[\hat{\theta}]).$$

Recalling that $\sigma_{\theta}/\sigma_{\hat{\theta}} = \rho_{\theta,\hat{\theta}}$ and reintroducing the subscript G to denote separate regression equations for $G = R$ and $G = F$, we obtain

$$E_G[\theta|\hat{\theta}] = E_G[\theta] + \rho_G^2(\hat{\theta} - E_G[\hat{\theta}]), \tag{7}$$

where ρ_G has been introduced to stand for $\rho_{\theta,\hat{\theta}}$ in group G. All the quantities on the right-hand side of the above equation can be estimated from data. Both $E_G[\theta]$ and $E_G[\hat{\theta}]$ are estimated by $\bar{\hat{\theta}}_G$, the mean value of $\hat{\theta}$ in group G. To obtain an estimate for ρ_G, an estimate of σ_e^2 in each group must be found. This is obtained by using the asymptotic relationship between σ_e^2 and the test information function, $I(\theta)$,

$$\sigma_e^2 = \frac{1}{E_G[I(\theta)]}. \tag{8}$$

An estimate of $E_G[I(\theta)]$ is obtained by estimating the test information for each group G test taker who took the studied item and then taking the average over all these test takers. The test information for a test taker is estimated based on the value of $\hat{\theta}$ for the test taker and on the values of the item parameters for the items that were used to obtain $\hat{\theta}$. For convenience, the true item parameters were used with each test taker's $\hat{\theta}$ estimate to obtain test information.

We now turn to estimation of β. In the paper-and-pencil setting, SIBTEST estimates β by matching R and F test takers on estimated true score, estimating their proportion-right scores on the studied item at each value of estimated true score, taking the difference between the R and F proportion-right scores at each value of estimated true score, and taking a weighted average of these differences over the different values of estimated true score, weighting by the total number of test takers at each value of estimated true score. Ideally, CATSIB would operate in exactly the same way except that test takers would be matched on $\hat{\theta}^*$ (our notation for $E[\theta|\hat{\theta}]$) instead of on estimated true score. Thus, ideally, the formula for $\hat{\beta}$, the estimator for β, would be

$$\hat{\beta} = \sum_{\hat{\theta}^* = \hat{\theta}^*_{min}}^{\hat{\theta}^*_{max}} [\hat{P}_R(\hat{\theta}^*) - \hat{P}_F(\hat{\theta}^*)]\hat{p}(\hat{\theta}^*), \tag{9}$$

where $P_G(\hat{\theta}^*)$ is the observed proportion of group G test takers with ability estimate $\hat{\theta}^*$ who got the studied item right and $\hat{p}(\hat{\theta}^*)$ is the observed proportion of R and F test takers at $\hat{\theta}^*$. Unfortunately, because $\hat{\theta}^*$ is a real-valued variable, only rarely do any two test takers have the exact same value of $\hat{\theta}^*$. Thus, it is impossible to calculate the proportion-right score on the studied item at a specific exact value of $\hat{\theta}^*$. To avoid this problem, the approach taken in this study was to divide the observed $\hat{\theta}^*$ range into n equal intervals. Test takers were then classified into one of the n intervals based on their values of $\hat{\theta}^*$. Hence, based on these intervals, $\hat{\beta}$ was calculated from the following equation,

$$\hat{\beta} = \sum_{k=1}^{n} [\hat{P}_{R,k} - \hat{P}_{F,k}]\hat{p}_k, \tag{10}$$

where $\hat{P}_{G,k}$ is the observed proportion of group G test takers in ability interval k who got the studied item right, and $\hat{p}_k$ is the observed proportion of R and F test takers who were classified into interval k.

Because the number of intervals was arbitrary, the approach we took was to have the computer program automatically determine the number of intervals. To ensure stable statistical estimation, an interval was required to have a minimum of three test takers from each of R and F for that interval to be included in the calculation of $\hat{\beta}$. All intervals with fewer than this minimum number were not used. Thus, it was important to carefully choose the number of intervals. If too many intervals were to be used, the intervals could become so sparsely populated with test takers that too many intervals (and, thus, too many test takers) could be eliminated from the statistic calculation resulting in a powerless statistic. On the other hand, if too few

intervals were to be used, the test statistic could become overly sensitive to impact and its Type I error could become unacceptably inflated. (In the extreme case of a single interval, the statistic would reduce to being purely a measure of impact.) To strike a balance between these two extremes, CATSIB was programmed to automatically start with an arbitrarily large number of ability intervals (80) and to then monitor how many test takers would be eliminated due to the throwing out of sparse cells. If more than 7.5% of either the R or F test takers would be eliminated, CATSIB will automatically decrease the number of cells until the number of test takers eliminated from each group becomes less than or equal to 7.5%. However, the minimum number of ability intervals was set at 20, even if this meant that the number of test takers eliminated from one or both of the groups sometimes exceeded 7.5%.

The standard error for $\hat{\beta}$ can then be estimated based on the observed variance of the studied item responses in each ability interval:

$$\hat{\sigma}(\hat{\beta}) = \sqrt{\sum_{k=1}^{n} \left[\frac{\hat{\sigma}^2_{R,k}(Y)}{n_{R,k}} + \frac{\hat{\sigma}^2_{F,k}(Y)}{n_{F,k}} \right] \hat{p}^2_k}, \tag{11}$$

where Y denotes the response to the studied item, $\sigma^2_{G,k}(Y)$ is the observed variance of Y in ability interval k for group G, and $n_{G,k}$ is the number of G group test takers in interval k.

The test statistic for testing the null hypothesis of no DIF is then given by

$$B = \frac{\hat{\beta}}{\hat{\sigma}(\hat{\beta})}. \tag{12}$$

The null hypothesis of no DIF is rejected at level α if the statistic B exceeds the $100(1-\alpha)$ percentile obtained from the standard normal table.

The estimate $\hat{\beta}$ serves as an index of the amount of DIF present in the item. For example, it is possible that an item may exhibit statistically significant DIF but the degree of DIF may not be practically meaningful in terms of how it affects the performance of test takers in the two groups. $\hat{\beta}$ can be very useful in assessing the degree of DIF practically. It can be seen from Equation 1 that $\hat{\beta}$ estimates the average difference between R and F groups in percent chance of a correct response (conditional on $\hat{\theta}^*$) on the studied item. When $\hat{\beta} = .050$, for example, the percent chance of a correct response on the studied item (conditional on ability) is estimated to be 5 percentage points higher for the reference group than for the focal group, which is classified here as moderate DIF. When $\hat{\beta} = .100$, on the other hand, the conditional percent chance of getting the studied item right is 10 percentage points higher for the reference group than for the focal group, which is classified here as large DIF. Throughout this paper we employ the DIF classification scheme suggested by Dorans (1989) in which $.050 \leq \hat{\beta} < .100$ is considered to indicate a moderate DIF item, and $\hat{\beta} \geq .100$ is considered to indicate a large DIF item. In other settings this categorization could also be subjective depending upon the nature of item parameters. For example, a $\hat{\beta} = .050$ might be considered a larger effect size for a difficult item than for an easy item because it is a greater percent of the maximum observed value.

Unique features of CATSIB are that it has a theoretically based correction for adjusting the $\hat{\theta}$s of the R and F groups to account for the effects of impact, and it has the capability of assessing DIF for either a single item or a collection of items.

The Simulation Study

At LSAC, as well as at other testing companies, DIF analyses are routinely carried out during the pretesting process. The simulation study will, therefore, mimic a pretest scenario in which test takers are adaptively administered a fixed-length CAT composed of operational items with well estimated item parameters and are linearly administered a certain number of pretest items with unknown statistical properties. The objectives of the proposed study are to assess the Type I error rate and power level of CATSIB for detecting DIF in these pretest items, using adjusted ability estimates from the adaptively administered operational CAT items as the matching criterion.

The length of the CAT was fixed at 25 items, which is typical of a standard CAT. That is, each simulated test taker was administered 25 items adaptively from a pool of 1,000 items. Additionally, all the simulated test takers were also linearly administered the same 16 pretest items that were to be evaluated for DIF. Three factors were varied in this study: the sample size, the impact level (d_T), and the amount of DIF (β). Three different combinations of test taker sample sizes were selected: $n_R = 250$, $n_F = 250$; $n_R = 500$, $n_F = 250$; and $n_R = 500$, $n_F = 500$; where n_R and n_F denote the sample sizes in the R and F groups, respectively. Three different impact levels were used: 0, .5, and 1. These three d_T levels correspond to differences in means of the ability

BEST COPY AVAILABLE

distributions of R and F groups that are 0, .5, and 1.0 standard deviations apart. The sample sizes and impact levels were completely crossed resulting in nine combinations of sample size and d_T level. Three DIF levels were used in the pretest items: no DIF ($\beta = 0$), moderate DIF ($\beta = .05$), and large DIF ($\beta = .10$).

The means of ability distributions for the R and F groups were determined in such a way that their weighted average was equal to the mean difficulty level of the CAT pool (which was equal to 0) and their difference was equal to the desired impact level. This was accomplished by solving for $\mu_{\theta R}$ and $\mu_{\theta F}$ from the following two equations:

$$\alpha_R \mu_{\theta R} + \alpha_F \mu_{\theta F} = 0$$

and

$$\mu_{\theta R} - \mu_{\theta F} = d_T$$

where

$$\alpha_R = \frac{n_R}{n_r + n_F}; \alpha_F = \frac{n_F}{n_r + n_F}.$$

The standard deviations of ability distributions were each set equal to 1.

In generating the item parameters of the item pool for the matching subtest, the goal was to generate parameters that closely resembled those estimated from real data. Upon observing the descriptive properties of the LSAT item pools, it was found that the distribution of item discrimination parameters ranged from .5 to 1.7 and followed a positively skewed distribution, while item difficulty parameters ranged from –2 to 2 and followed the standard normal distribution. The discrimination parameters were therefore generated from a lognormal distribution, and difficulty parameters were generated from the standard normal distribution. The lower asymptote was independently generated from a uniform distribution to range between .12 and .22 to approximate those from actual LSAT data. The precise distributions used for the item parameters are described below:

$$log(a) \sim \text{normal } (-.357, .25) \text{ for } b \leq -1 \text{ with range } .4 \leq a \leq 1.1$$

$$log(a) \sim \text{normal } (-.223, .34) \text{ for } b > -1 \text{ with range } .4 \leq a \leq 1.7$$

$$b \sim N(0, 1) \text{ with range } -3 \leq b \leq 3$$

$$c \sim U(.12, .22).$$

DIF was introduced in the pretest studied items through differences in the difficulty parameters between the R and F groups using the following model for DIF:

$$\text{DIF} = \hat{\beta} = \int [P_R(\theta) - P_F(\theta)] f(\theta) d\theta, \tag{13}$$

where

$$P_G(\theta) = c + \frac{1-c}{1 + exp[-1.7a(\theta - b_G)]}, \; G = R \text{ or } F. \tag{14}$$

There were in total 16 DIF items: 6 with $\beta = 0$, 5 with $\beta = .05$, and 5 with $\beta = .1$. For $\beta = 0$, six types of studied items were chosen. These items are arbitrarily labeled item 1 to 6 and have the following respective a and b parameters: (.4, –1.5), (.4,1.5), (.8, 0), (1.0, –1.5), (1.4, 1.5), and (1.4, –1.5). The item with medium

discrimination (.8) and medium difficulty (0) was included to represent an average LSAT item. Four of the items represented somewhat extreme combinations of a and b as observed in LSAT data: low discrimination (.4) and low difficulty (–1.5); low discrimination (.4) and high difficulty (1.5); moderately high discrimination (1.0) and low difficulty (–1.5); and high discrimination (1.4) and high difficulty (1.5). One more extreme item (which generally does not occur with LSAT data) with high discrimination (1.4) and low difficulty (–1.5) was included among the $\beta = 0$ items because previous studies had shown that high discriminating easy items have a tendency for impact-induced Type I error inflation (Allen & Donoghue, 1996; Roussos & Stout, 1996).

The a and b parameters for the 5 items with $\beta = .050$ and the 5 with $\beta = .100$ were chosen to be parallel to the item parameters that were assigned to the first 5 items with $\beta = 0$. The a parameters of items 7 to 11 and 12 to 16 were set exactly equal to the a parameters of items 1 to 5. Because items 7 to 16 had non-zero DIF, we could not set b_R and b_F equal to the b parameters of items 1 to 5. To keep the b parameters of items 7 to 11 and 12 to 16 similar to those of items 1 to 5, we required the weighted average of b_R and b_F (weighted by n_R and n_F) for items 7 to 11 and 12 to 16 to be equal to the b parameters for items 1 to 5, respectively. The precise values for b_R and b_F for items 7 to 11 and 12 to 16 are listed in Table 1.

TABLE 1
Item parameters for the simulated studied items with $\beta = .050$ and $\beta = .100$

			Difference in R and F Ability Means (d_T)					
			$d_T = 0$		$d_T = .5$		$d_T = 1$	
Item	β	a	b_R	b_F	b_R	b_F	b_R	b_F
Equal Sizes for Reference Group and Focal Group								
7	0.050	0.4	-1.738	-1.262	-1.739	-1.261	-1.741	-1.259
8	0.050	0.4	1.262	1.738	1.261	1.739	1.259	1.741
9	0.050	0.8	-0.120	0.120	-0.122	0.122	-0.129	0.127
10	0.050	1.0	-1.691	-1.309	-1.691	-1.309	-1.689	-1.311
11	0.050	1.4	1.303	1.697	1.305	1.695	1.310	1.690
12	0.100	0.4	-1.977	-1.023	-1.978	-1.022	-1.983	-1.017
13	0.100	0.4	1.023	1.977	1.022	1.978	1.017	1.983
14	0.100	0.8	-0.241	0.241	-0.244	0.244	-0.254	0.254
15	0.100	1.0	-1.882	-1.118	-1.881	-1.119	-1.878	-1.122
16	0.100	1.4	1.109	1.891	1.113	1.887	1.122	1.878
Reference Group Twice as Large as Focal Group								
7	0.050	0.4	-1.656	-1.188	-1.656	-1.118	-1.657	-1.186
8	0.050	0.4	1.338	1.824	1.338	1.824	1.337	1.826
9	0.050	0.8	-0.080	0.160	-0.081	0.162	-0.084	0.168
10	0.050	1.0	-1.622	-1.256	-1.622	-1.256	-1.622	-1.256
11	0.050	1.4	1.359	1.782	1.361	1.778	1.367	1.766
12	0.100	0.4	-1.806	-0.888	-1.807	-0.886	-1.810	-0.880
13	0.100	0.4	1.168	2.164	1.167	2.166	1.166	2.168
14	0.100	0.8	-0.161	0.322	-0.163	0.326	-0.168	0.336
15	0.100	1.0	-1.734	-1.032	-1.734	-1.032	-1.736	-1.028
16	0.100	1.4	1.198	2.104	1.204	2.092	1.217	2.066

Items 1 to 6 were used for determining the Type I error rate of CATSIB. Items 6 to 11 and 12 to 16 were used to investigate the power performance of CATSIB. All the pretest studied items had c parameters equal to 0.17, the average estimated c parameter for the LSAT data on which our item parameters were based.

The CAT Procedure

For a given combination of test taker sample sizes (n_R and n_F) and impact level (d_T), test takers of R and F groups were simulated from their respective distributions[1]. Each test taker in each of the groups R and F was adaptively administered a fixed length test of 25 items from a pool of 1,000 operational items. The ability estimates of test takers were determined using a standard maximum-information CAT design described as follows.

[1]The R Group test takers were simulated from $N(\mu_{\theta R,1})$ distribution and the F group test takers were simulated from $N(\mu_{\theta F,1})$ distribution.

BEST COPY AVAILABLE

The ability scale from –2.25 to 2.25 was divided into 37 equal intervals in increments of 0.125. For each item i, item information, $I_i(\theta)$, was computed at the θ values corresponding to the midpoints of the 37 intervals using the following formula (Hambleton, Swaminathan, & Rogers, 1991, p.91):

$$I_i(\theta) = \frac{(1.7a_i)^2(1 - c_i)}{[c_i + exp(1.7a_i(\theta - b_i))][1 + exp(-1.7a_i(\theta - b_i))]^2},$$

where a_i, b_i, and c_i denote discrimination, difficulty, and lower asymptote parameters of item i respectively. At each θ level the pool of operational items was sorted according to the item information values from lowest to highest and saved in a separate table. This table was used during the simulations to select items with the highest information at a given θ level.

To prevent items from becoming overexposed, an exposure control method was incorporated (Kingsbury & Zara, 1989). Accordingly, the first item to be administered to a simulated test taker was randomly selected from the 10 items with highest information values at $\theta = 0$ (the starting value for all simulated test takers). The second item was randomly selected from the 9 best items at the new estimate of θ. The third item was randomly selected from the 8 best items, and so on until, beginning with the 10th item, the item with the highest information was selected (unless, of course, the item had already been administered to that simulated test taker, in which case the next best item was selected).

After administering each item in this manner to each test taker, the simulated test taker's response (right/wrong) was determined, and the simulated test taker's estimated ability, $\hat{\theta}$, was updated using Owen's Bayesian sequential scoring (Owen, 1969). After all 25 items were administered, a Bayesian modal score was calculated and was used as the final ability estimate ($\hat{\theta}$).

The DIF items were then administered nonadaptively one at a time to all the test takers. After each administration of a DIF item, the DIF estimate $\hat{\beta}$ and the statistic B were computed and tested for the presence of DIF using a right-tailed test, a left-tailed test, and a two-tailed test. The left-tailed test involves rejecting the null hypothesis of no DIF at level $\alpha = .05$ if the computed z-statistic is less than –1.645. The right-tailed test rejects if the computed z-statistic is greater than 1.645; and the two-tailed test rejects if $|z| \geq 1.96$. For a given combination of sample size and the d_T level, this process, starting from the simulation of θs, was replicated 400 times. The average DIF estimate $\bar{\hat{\beta}}$ over 400 replications, its standard error, the rejection rates for the right-tailed test, the left-tailed test, and the two-tailed test were computed for CATSIB with the regression correction (CATSIB WRC) and for CATSIB without the regression correction (CATSIB WORC).

Results

The Type I Error Study

The results of the Type I error study are reported in Tables 2 through 5 and in Figure 1 for the no-DIF ($\beta = 0$) items (1 to 6). Table 2 shows the DIF estimation results: the mean values for $\hat{\beta}$ ($\bar{\hat{\beta}}$) over the 400 trials, along with the standard errors for these mean values. The estimated DIF was close to 0 for CATSIB WRC for all the studied items, even for high levels of impact. On the other hand, CATSIB WORC displayed increasingly biased estimation of β as impact increased. As expected, the standard errors of $\hat{\beta}$ are similar for CATSIB WRC and CATSIB WORC and show the expected behavior of decreasing with increasing sample size.

TABLE 2

Type I error study DIF estimation results tabulated values: $\hat{\beta}$ (and its standard error)

	CATSIB WRC			CATSIB WORC		
	Sample Sizes for Reference/Focal Groups					
Item	250/250	500/250	500/500	250/250	500/250	500/500
	No Group Differences in Ability Means (d_T = 0)					
1	.001 (.0018)	-.002 (.0016)	.000 (.0014)	.001 (.0018)	-.002 (.0016)	.000 (.0014)
2	-.003 (.0023)	-.002 (.0020)	.001 (.0015)	-.003 (.0023)	-.002 (.0020)	.001 (.0016)
3	-.001 (.0022)	.001 (.0019)	-.001 (.0015)	.000 (.0021)	.001 (.0019)	.000 (.0015)
4	-.001 (.0014)	.000 (.0012)	-.001 (.0010)	-.001 (.0014)	.000 (.0012)	-.001 (.0010)
5	.004 (.0019)	-.003 (.0015)	-.001 (.0014)	.004 (.0019)	-.003 (.0015)	-.001 (.0014)
6	.002 (.0012)	.000 (.0010)	.000 (.0008)	.002 (.0012)	.000 (.0010)	.000 (.0008)
	Half Standard Deviation Difference in Group Ability Means (d_T = .5)					
1	.000 (.0019)	-.002 (.0017)	.000 (.0015)	.003 (.0018)	.000 (.0017)	.003 (.0015)
2	-.002 (.0022)	-.001 (.0020)	.003 (.0016)	.000 (.0022)	.002 (.0020)	.006 (.0015)
3	.000 (.0021)	.001 (.0018)	-.001 (.0016)	.005 (.0021)	.007 (.0018)	.005 (.0015)
4	.000 (.0015)	.002 (.0012)	-.001 (.0010)	.003 (.0015)	.005 (.0012)	.002 (.0010)
5	.002 (.0019)	-.004 (.0017)	-.001 (.0014)	.005 (.0019)	.000 (.0017)	.003 (.0014)
6	.001 (.0013)	.002 (.0009)	.002 (.0009)	.004 (.0013)	.005 (.0010)	.004 (.0009)
	One Standard Deviation Difference in Group Ability Means (d_T = 1)					
1	.001 (.0021)	-.001 (.0019)	.001 (.0016)	.007 (.0020)	.006 (.0019)	.007 (.0016)
2	.000 (.0025)	-.001 (.0022)	.004 (.0018)	.005 (.0025)	.006 (.0022)	.009 (.0017)
3	-.004 (.0024)	.000 (.0022)	.000 (.0016)	.008 (.0023)	.012 (.0021)	.011 (.0016)
4	.003 (.0017)	.004 (.0014)	.002 (.0012)	.009 (.0016)	.010 (.0014)	.008 (.0012)
5	.000 (.0021)	-.005 (.0019)	-.002 (.0015)	.008 (.0021)	.003 (.0019)	.007 (.0015)
6	.001 (.0015)	.004 (.0011)	.003 (.0010)	.008 (.0014)	.010 (.0011)	.009 (.0010)

Tables 3 and 4 show the rejection rate results for CATSIB WRC and CATSIB WORC, respectively. Because α = .05, the observed rejection rates in Tables 3 and 4 would be expected to fall between .03 and .07, 95% of the time (based on the exact binomial distribution) if the procedures were adhering well to the nominal level of .05. It can be seen from these tables that some of the rejection rates were out of bounds, below .03 or above .07. Figure 1 graphically displays the number of rejection rates that were out of bounds by d_T (impact) level for both CATSIB WRC and CATSIB WORC. Each plotted point is the number of rejection rates out of bounds out of 54 cases (6 items x 3 hypothesis tests x 3 sample sizes). Based on exact binomial probabilities, out of 54 tests, one expects 0 to 5 out of bounds due to chance alone about 95% of the time. From Figure 1 it can be seen for CATSIB WORC that, as the level of impact increases, the number of tests out of bounds also increases. It is evident that as the impact level increases, the Type I error inflation also increases, but the degree of inflation is much steeper for CATSIB without regression correction. For the large impact level (d_T = 1.0), the number of rejection rates out of bounds for CATSIB WORC were grossly inflated over the chance levels, while for CATSIB WRC the inflation was only slightly more than the chance level.

BEST COPY AVAILABLE

TABLE 3
Type I error study rejection rate results for CATSIB WRC

	Sample Sizes for Reference/Focal Groups								
	250/250			500/250			500/500		
Item	Left-Tailed	Right-Tailed	Two-Tailed	Left-Tailed	Right-Tailed	Two-Tailed	Left-Tailed	Right-Tailed	Two-Tailed
	No Differences in Ability Means ($d_\tau = 0$)								
1	0.0275	0.0500	0.0425	0.0675	0.0400	0.0450	0.0625	0.0725	0.0725
2	0.0500	0.0425	0.0525	0.0550	0.0500	0.0600	0.0525	0.0575	0.0500
3	0.0600	0.0575	0.0475	0.0425	0.0500	0.0475	0.0575	0.0425	0.0650
4	0.0325	0.0475	0.0425	0.0525	0.0550	0.0500	0.0525	0.0450	0.0625
5	0.0375	0.0575	0.0500	0.0450	0.0300	0.0250	0.0500	0.0475	0.0450
6	0.0575	0.0550	0.0700	0.0625	0.0300	0.0375	0.0425	0.0400	0.0450
	Half Standard Deviation Difference in Group Ability Means ($d_\tau = .5$)								
1	0.0375	0.0600	0.0325	0.0600	0.0400	0.0475	0.0625	0.0550	0.0625
2	0.0475	0.0350	0.0475	0.0475	0.0500	0.0425	0.0350	0.0525	0.0450
3	0.0600	0.0425	0.0375	0.0400	0.0500	0.0400	0.0575	0.0425	0.0450
4	0.0450	0.0700	0.0650	0.0500	0.0675	0.0575	0.0600	0.0500	0.0525
5	0.0425	0.0325	0.0325	0.0375	0.0475	0.0400	0.0475	0.0550	0.0425
6	0.0475	0.0575	0.0600	0.0475	0.0400	0.0425	0.0400	0.0750	0.0600
	One Standard Deviation Difference in Group Ability Means ($d_\tau = 1$)								
1	0.0300	0.0550	0.0275	0.0650	0.0375	0.0450	0.0600	0.0650	0.0650
2	0.0375	0.0500	0.0425	0.0500	0.0350	0.0350	0.0450	0.0375	0.0350
3	0.0450	0.0325	0.0500	0.0750	0.0500	0.0750	0.0400	0.0400	0.0475
4	0.0325	0.0775	0.0650	0.0500	0.0875	0.0800	0.0600	0.0700	0.0750
5	0.0425	0.0325	0.0350	0.0500	0.0575	0.0500	0.0525	0.0375	0.0650
6	0.0525	0.0725	0.0775	0.0400	0.0725	0.0550	0.0275	0.0950	0.0625

TABLE 4
Type I error study rejection rate results for CATSIB WORC

	Sample Sizes for Reference/Focal Groups								
	250/250			500/250			500/500		
Item	Left-Tailed	Right-Tailed	Two-Tailed	Left-Tailed	Right-Tailed	Two-Tailed	Left-Tailed	Right-Tailed	Two-Tailed
	No Differences in Ability Means ($d_\tau = 0$)								
1	0.0275	0.0525	0.0425	0.0675	0.0350	0.0475	0.0650	0.0700	0.0725
2	0.0550	0.0425	0.0500	0.0550	0.0575	0.0625	0.0500	0.0625	0.0525
3	0.0575	0.0450	0.0525	0.0400	0.0550	0.0375	0.0600	0.0425	0.0675
4	0.0425	0.0475	0.0400	0.0625	0.0575	0.0525	0.0625	0.0525	0.0600
5	0.0425	0.0600	0.0525	0.0375	0.0300	0.0275	0.0525	0.0450	0.0425
6	0.0600	0.0525	0.0675	0.0650	0.0350	0.0375	0.0475	0.0525	0.0525
	Half Standard Deviation Difference in Group Ability Means ($d_\tau = .5$)								
1	0.0375	0.0700	0.0375	0.0550	0.0500	0.0475	0.0450	0.0700	0.0675
2	0.0425	0.0400	0.0425	0.0375	0.0550	0.0575	0.0200	0.0575	0.0475
3	0.0450	0.0650	0.0375	0.0275	0.0675	0.0325	0.0250	0.0700	0.0550
4	0.0425	0.0825	0.0600	0.0325	0.0850	0.0675	0.0450	0.0725	0.0550
5	0.0350	0.0450	0.0325	0.0425	0.0475	0.0425	0.0350	0.0700	0.0400
6	0.0350	0.0750	0.0650	0.0325	0.0600	0.0450	0.0325	0.1025	0.0575
	One Standard Deviation Difference in Group Ability Means ($d_\tau = 1$)								
1	0.0275	0.0775	0.0375	0.0425	0.0475	0.0500	0.0325	0.0775	0.0700
2	0.0350	0.0525	0.0375	0.0225	0.0500	0.0325	0.0325	0.0700	0.0425
3	0.0300	0.0725	0.0375	0.0375	0.1000	0.0750	0.0200	0.0850	0.0600
4	0.0250	0.1025	0.0700	0.0275	0.1500	0.0900	0.0275	0.1125	0.0900
5	0.0200	0.0650	0.0425	0.0375	0.0775	0.0525	0.0300	0.0800	0.0600
6	0.0300	0.1050	0.0800	0.0200	0.1325	0.0750	0.0150	0.1500	0.1050

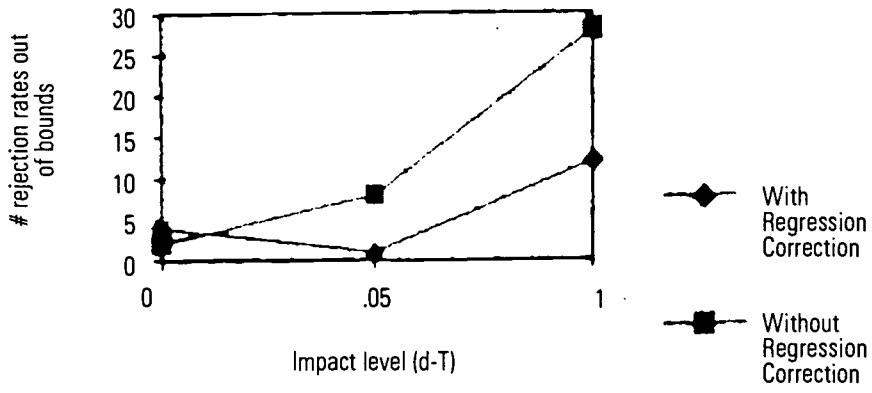

FIGURE 1. *Type I error study: Number of rejection rates out of bounds*

The summaries of Type I error rates by sample size and d_T level are reported in Table 5. For this table the observed rejection rates are expected to fall between .0413 and .0587, 95% of the time if the procedures are adhering well to the nominal .05 rejection rate. It can be seen from Table 5 that, while the observed average rejection rates for CATSIB WRC were generally within chance of the nominal .05 level, they were either significantly inflated or deflated for CATSIB WORC in the presence of impact with the majority (75%) of them being out of bounds. For CATSIB WORC there was a consistent pattern of serious deflation of Type I error rates for the left-tailed hypothesis test across all levels of sample size, and a corresponding serious inflation of Type I error rates for the right-tailed hypothesis test. Averaging across all items, sample sizes, and impact levels, at the nominal .05 level the observed rejection rates would be expected to fall between .0471 and .0529, 95% of the time. The CATSIB WRC Type I error rates were .0487, .0518, and .0509 for the left-, right-, and two-tailed hypothesis tests, respectively, which were all within the expected bounds. The corresponding CATSIB WORC Type I error rates were .0394, .0683, and .0540, which were all out of bounds.

TABLE 5
Summary of Type I error rejection rate results

	With Regression Correction			Without Regression Correction		
Sample Sizes	Difference in R and F Ability Means (d_T)					
for R/F	$d_T = 0$	$d_T = .5$	$d_T = 1$	$d_T = 0$	$d_T = .5$	$d_T = 1$
Left-Tailed Rejection Rates						
250/250	0.0442	0.0467	0.0400	0.0475	0.0396	0.0279
500/250	0.0542	0.0471	0.0550	0.0546	0.0379	0.0313
500/500	0.0529	0.0504	0.0475	0.0563	0.0338	0.0263
mean	0.0504	0.0481	0.0475	0.0528	0.0371	0.0285
Right-Tailed Rejection Rates						
250/250	0.0517	0.0496	0.0533	0.0500	0.0629	0.0792
500/250	0.0425	0.0492	0.0567	0.0450	0.0608	0.0929
500/500	0.0508	0.0550	0.0575	0.0542	0.0738	0.0958
mean	0.0483	0.0513	0.0558	0.0497	0.0658	0.0893
Two-Tailed Rejection Rates						
250/250	0.0508	0.0458	0.0496	0.0508	0.0458	0.0508
500/250	0.0442	0.0450	0.0567	0.0442	0.0488	0.0625
500/500	0.0567	0.0513	0.0583	0.0579	0.0538	0.0713
mean	0.0506	0.0474	0.0549	0.0510	0.0494	0.0615

BEST COPY AVAILABLE

The Power Study

For this study, only the CATSIB WRC results will be discussed because the CATSIB WORC results were distorted by the statistical bias evident in its Type I error results (see Table 2). The power study estimation and rejection rate results are presented in Table 6. In Table 6, items 7 to 11 are of moderate DIF (β = .050) and 12 to 16 are of large DIF (β = .100). Within each group, items are mixed in terms of low and high difficulty and discrimination parameters (see Table 1). Since only positive values of β were induced (DIF against the focal group), only right-tailed and two-tailed rejection rates were of interest, and only the two-tailed results are presented in Table 6.

TABLE 6

Power study estimation and two-tailed rejection rate (2trr) results for CATSIB WRC

	β = .050						β = .100					
	Sample Sizes for Reference/Focal Groups						Sample Sizes for Reference/Focal Groups					
	250/250		500/250		500/500		250/250		500/250		500/500	
Item	$\hat{\beta}$	2Trr	$\hat{\beta}$	2Trr	$\hat{\beta}$	2Trr	$\hat{\beta}$	2Trr	$\hat{\beta}$	2Trr	$\hat{\beta}$	2Trr
	No Differences in Ability Means (d_T = 0)											
7	0.0470	0.2400	0.0520	0.3200	0.0500	0.4450	12	0.0980	0.7200	0.1010	0.8375	0.1030	0.9625
8	0.0500	0.2400	0.0480	0.2625	0.0500	0.3425	13	0.0980	0.5775	0.0990	0.7125	0.1010	0.8750
9	0.0490	0.2100	0.0540	0.3100	0.0520	0.4025	14	0.1010	0.6875	0.1050	0.8100	0.1020	0.9250
10	0.0490	0.4525	0.0470	0.4725	0.0500	0.7375	15	0.1000	0.9500	0.1020	0.9725	0.1000	1.0000
11	0.0480	0.2325	0.0490	0.3150	0.0500	0.4600	16	0.0970	0.7025	0.1000	0.8650	0.0980	0.9350
mean	0.0490	0.2750	0.0500	0.3360	0.0500	0.4775		0.0990	0.7275	0.1010	0.8395	0.1010	0.9395
	Half Standard Deviation Difference in Group Ability Means (d_T = .5)											
7	0.0480	0.2300	0.0530	0.3200	0.0500	0.4300	12	0.1000	0.7125	0.1040	0.8250	0.1020	0.9450
8	0.0500	0.2100	0.0480	0.2400	0.0480	0.3250	13	0.0990	0.5625	0.0970	0.6925	0.1020	0.8750
9	0.0500	0.2375	0.0530	0.2925	0.0520	0.4000	14	0.1040	0.6775	0.1070	0.8025	0.1030	0.9250
10	0.0500	0.3975	0.0500	0.5125	0.0500	0.7300	15	0.1020	0.9525	0.1050	0.9850	0.1000	1.0000
11	0.0460	0.2225	0.0460	0.2725	0.0500	0.4225	16	0.0970	0.6675	0.0920	0.7475	0.0970	0.9275
mean	0.0490	0.2595	0.0500	0.3275	0.0500	0.4615		0.1000	0.7145	0.1010	0.8105	0.1010	0.9345
	One Standard Deviation Difference in Group Ability Means (d_T = 1)											
7	0.0500	0.2275	0.0530	0.2850	0.0510	0.3700	12	0.1010	0.6100	0.1070	0.7475	0.1030	0.9150
8	0.0520	0.1725	0.0470	0.2025	0.0490	0.2650	13	0.1010	0.4725	0.1000	0.6275	0.1030	0.8350
9	0.0510	0.2075	0.0530	0.2300	0.0520	0.3625	14	0.1090	0.6250	0.1080	0.6950	0.1070	0.8825
10	0.0520	0.3850	0.0550	0.5400	0.0520	0.6350	15	0.1010	0.8875	0.1120	0.9800	0.1010	0.9925
11	0.0430	0.1825	0.0400	0.2125	0.0500	0.3525	16	0.0940	0.5925	0.0800	0.5425	0.0960	0.8450
mean	0.0500	0.2350	0.0500	0.2940	0.0510	0.3970		0.1010	0.6375	0.1010	0.7185	0.1020	0.8940

In terms of DIF estimation, in almost all cases, the average amount of estimated DIF ($\hat{\beta}$) was close to the true values (.050 for items 7 to 11, and .100 for items 12 to 16). However, there were a few cases of estimation bias. For the highest level of impact and the smallest two sample sizes, the high discriminating, high difficulty level items (items 11 and 16) were consistently underestimated with the underestimation being fairly substantial for the case of unequal *R* and *F* sample sizes ($\hat{\beta}$ values of .040 and .080, respectively, for items 11 and 16). Also, an approximately 10% overestimation (i.e., positive) bias occurred with items 10 and 15 when impact was highest and the reference group size was twice the focal group size. To explore the reason for these statistical biases, we monitored the average number of ability interval cells used in the CATSIB statistic calculations along with the average percentages of *R* and *F* test takers included in those cells. These results are presented in Table 7. At all three levels of sample size for the two lowest levels of impact (d_T = 0 and .5) and at the highest level of sample size for d_T = 1, CATSIB generally achieved on average the goal of including 92.5% or more of the *R* and *F* test takers in the statistic calculation. However, at d_T = 1 at the lowest two sample sizes, the average percentage of test takers included for *R* in the 500/250 case and for both *R* and *F* in the 250/250 case was well below the targeted 92.5%. As indicated in Table 7, the reason for this was that CATSIB was constrained to use no fewer than 20 ability interval cells while fewer than 20 cells were needed for these cases. Thus, the automatic reducing of the number of ability interval cells experienced an unexpected floor effect. Hence, the underestimation of β for items 11 and 16 for d_T = 1 with the smallest sample sizes is probably due to the exclusion of too many test takers from the statistic calculation. Similarly, the overestimation bias with items 10 and 15 may be due to the reference group having its test takers excluded at a higher rate than the focal group. Lowering the limit on the minimum number of cells might eliminate this bias. The biases almost totally disappear at the highest sample size.

TABLE 7
Average number of cells used and percentage of test takers included in CATSIB statistic calculation

Sample Sizes for R/F	Average Number of Cells Used			Average Percentage of Test Takers Included					
				Reference Group			Focal Group		
	Difference in R and F Ability Means(d_T)								
	$d_T = 0$	$d_T = .5$	$d_T = 1$	$d_T = 0$	$d_T = .5$	$d_T = 1$	$d_T = 0$	$d_T = .5$	$d_T = 1$
250/250	29.3	22.0	20.0	94.7	93.2	85.6	94.7	93.9	87.6
500/250	36.8	26.2	20.1	94.0	93.3	86.4	95.4	95.9	93.1
500/500	63.2	44.5	21.2	94.2	94.2	92.3	94.1	94.4	93.0

We now turn to the rejection rate results in Table 6. The detection of large DIF (β values of .100 or more) at the pretest stage is a critical requirement of a CAT DIF procedure. The results in Table 6 show that CATSIB has power rates of over 90% for 10 out of 15 cases for the $\beta = .100$ items for sample sizes of 500 in each group. And even when sample size is as small as 250 in each group, the CATSIB power rates for the $\beta = .100$ items still were over 66.7% for 9 out of 15 of the $\beta = .100$ cases. Because large DIF is defined as $\beta \geq .100$, the detection rates for the $\beta = .100$ case represents the minimum CATSIB large DIF power rates. That is, the CATSIB power rates can be expected to be substantially higher for most large DIF items because most large DIF items will have β values larger than .100.

As expected, the rejection rates generally increased as the amount of DIF increased and as sample size increased. Averaging over items 7 to 11 and separately over items 12 to 16, as the DIF level increased from moderate ($\beta = .05$) to large ($\beta = .1$), the average power rates went up from 34% to 80% for the two-tailed hypothesis test. As sample size increased from 500 ($n_R = 250$, $n_F = 250$) to 1,000 ($n_R = 500$, $n_F = 500$), the average power for the 5 items with $\beta = .100$ increased from 69% to 92% for the two-tailed hypothesis test. Even for small samples such as $n_R = 250$, $n_F = 250$, the power was remarkably high (60% or more for most items) for $\beta = .100$.

The rates also varied across the different items. This variation is mostly characterized by rejection rates being generally higher for items that have higher discrimination and lower difficulty levels. The discrimination effect can probably be attributed to the higher discriminations reducing the error variance in the item response data. The difficulty level effect is probably due to the guessing parameter having a smaller effect on the item responses, which, again decreases the error variance in the data.

Impact level also had some effect on the rejection rates. Generally, as impact increased, the rejection rates went down. This effect was mostly constrained to going from $d_T = 0.5$ to $d_T = 1.0$. There was very little effect in going from $d_T = 0$ to $d_T = 0.5$. And for the DIF values of most interest, $\beta = .100$, the effect of impact was practically negligible for the largest sample size. In other words, given 500 test takers in each group, the power of CATSIB to detect any large DIF values is quite high and almost unaffected by impact.

In summary, the simulation results have shown that CATSIB with the regression correction has exhibited good statistical DIF detection properties. The regression correction was very effective in controlling for Type I error even for impact levels as large as one standard deviation.

The power rates for items exhibiting the minimum amount of DIF to be considered large DIF were exceptionally high (usually over 90%) when there were 500 test takers in each group and generally well over 60% when the number of test takers was as small as 250 in each group.

Concluding Remarks

This study has shown that CATSIB can be a practical and reliable statistical procedure for detecting DIF on computerized adaptive tests. It has performed satisfactorily under the conditions controlled for in this study and therefore shows high potential for operational use. The use of $\hat{\beta}$ for assessing the amount of DIF can be very useful in applications. Because $\hat{\beta}$, the estimated degree of DIF, is the difference in probabilities of correct responses between R and F groups on the studied item, this index can be used in judgments about whether or not to keep an item in the pool for future administrations.

Further studies are needed to investigate the performance of CATSIB. For the most difficult items in the presence of large impact, too many test takers were excluded from the statistic calculation because the minimum number of interval cells was fixed at 20. Future simulations are planned where the minimum number of cells is allowed to go as low as necessary to meet the required percentage of included test takers. As another way of dealing with the difficulty of matching reference and focal group test takers on difficult items in the presence of large impact, a kernel-smoothed version of CATSIB is also under development. Future studies are also planned to increase the realism in the simulation studies. Future studies will use estimated item parameters rather than the known true values; also the introduction of content constraints, other exposure control algorithms, and multistage testlet designs are being considered. Furthermore, CATSIB's performance on real data must also be evaluated.

References

Allen, N. L., & Donoghue, J. R. (1996). Applying the Mantel-Haenszel procedure to complex samples of items. *Journal of Educational Measurement, 33,* 231–251.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2,* 217–233.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355–368.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2,* 359–375.

Owen, R. J. (1969) *Bayesian approach to tailored testing* (Research Bulletin 69-92). Princeton, NJ: Educational Testing Service.

Roussos, L. A. (1996, June). *A type I error rate study of a modified SIBTEST DIF procedure with potential application to computerized-adaptive tests.* Paper presented at the annual meeting of the Psychometric Society, Banif, Alberta, Canada.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement, 33,* 215–230.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58,* 159–194.

Zwick, R., Thayer, D.T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement, 18,* 121–140.

Zwick, R., Thayer, D.T., & Wingersky, M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement, 32,* 341–363.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

X	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").